2023 IEEE 12th International Conference on Cloud Networking (CloudNet) | 979-8-3503-1306-2/23/$31.00 ©2023 IEEE | DOI: 10.1109/CloudNet59005.2023.10490081

Unveiling Equity: Exploring Feature Dependency using Complex-Valued Neural Networks and Attention Mechanism for Fair Data Analysis

1st Xuting Tang
Computer Science Department
Stevens Institute of Technology
Hoboken, USA
xtang18@stevens.edu

2nd Mengjiao Zhang
Computer Science Department
Stevens Institute of Technology
Hoboken, USA
mzhang49@stevens.edu

3rd Abdul Rafae Khan
Computer Science Department
Stevens Institute of Technology
Hoboken, USA
akhan4@stevens.edu

4th Steve Y. Yang
School of Business
Stevens Institute of Technology
Hoboken, USA
steve.yang@stevens.edu

5th Jia Xu
Computer Science Department
Stevens Institute of Technology
Hoboken, USA
jxu70@stevens.edu

***Abstract*—With the increasing use of big data, cloud computing, and machine learning in high-stake domains such as justice systems, financial institutions, and healthcare, concerns about fairness have become more prominent. This paper presents a novel approach to foster fair decision-making by tackling social bias and enhancing transparency in machine learning models. The proposed framework leverages quantum-inspired complex-valued neural networks and attention-based networks, offering improved transparency in modeling the decision process for interpreting feature importance and dependency. Furthermore, our approach tackles the challenges posed by imbalanced data through the incorporation of focal loss and oversampling techniques, resulting in reduced prediction errors. Through extensive experiments conducted on real-life datasets encompassing criminal charge prediction, financial fraud detection, and credit card default payment prediction, our approach consistently demonstrates reliable prediction precision and recall. Notably, our analysis of feature significance highlights the statistical importance of task-related features such as historical records of bank transactions or criminal charge history, while socially biased identifiers like race, gender, and age exhibit minimal significance. By excluding these biased features, our approach enhances fairness without compromising prediction accuracy, thereby contributing to the advancement of fair decision-making in big data and cloud computing across various high-stake domains.**

***Index Terms*—Fair Machine Learning in Cloud Computing, Model Interpretability, Data Analysis**

I. Introduction

Artificial intelligence (AI) has shown great success and potential when equipped with extensive data and the power of cloud computing. The AI systems can achieve and surpass human-level performance in various tasks such as image classification, speech recognition, language processing, and decision making. Furthermore, numerous high-stakes domains, such as finance [1], healthcare [2] and crime prediction [3] are harnessing the power of AI in conjunction with cloud computing. Despite the achievements of deep learning (DL) models and the increasingly critical role they play in these domains, AI models are considered “black-box” approaches and opaque to humans. The complex model architectures and many parameters make the models hard to understand. This DL black-box decision-making process is a severe problem for some sensitive domains like health care, criminal justice, finance, and other applications related to human life, rights, and privacy [4], [5]. One big issue is that the non-transparency problem often leads to ethical issues [5], [6]. For example, race, gender, and education level should not be the critical factors for DL decisions and should not cause algorithmic discrimination [7]. Therefore, besides high accuracy, we must also make our AI model transparent so humans can understand and control the decision-making process.

In this work, our goal is to increase the accuracy and understand the decision-making process of our prediction model. We investigate (1) what kind of information is influential in the final decision, and (2) how to identify unbiased features on the decision output. More specifically, our work presents a criminal charge prediction task and several financial prediction tasks and we introduce a framework that leads to an accurate decision with model interpretability.

The first technical challenge in our study is the heavily *imbalanced data*. For both the criminal charge prediction and the financial tasks, the datasets are highly imbalanced. The negative samples are ten times more numerous than the positive samples. If we use a deep learning model with standard cross-entropy loss, the prediction has very low recall and F-measure scores despite high precision. Because the positive samples are much fewer than the negative samples, the model tends to predict everything negative to minimize classification errors.

The second challenge is the explanation of the model predictions in the decision-making tasks. The crime charge prediction task is based on the criminal history of a person. The prediction results of a criminal charge or financial risk should be determined by the crime history and financial history instead of a person's demographic characteristics like race, age, and gender, which are biases for the model. However, existing prediction models on these tasks do not consider the interpretability [8]–[11]. It is unknown what feature types the models focus more on, a problem that can lead to ethical issues like gender discrimination.

To address the heavily imbalanced data, we adopt two methods. The first one is using focal loss instead of the normal cross-entropy loss to force the learning focus more on difficult examples with less similar training data to learn from. The second way is applying oversampling to make the datasets more balanced.

To explain the model prediction, we incorporate attention mechanisms in our models. We also adopt complex-valued neural networks (CVN), which can interpret the model with physical meanings. For interpreting with attention, we analyze feature importance using values of attention weights learned from the attentional Bi-LSTMs. We adopt additive attention to the Bi-LSTM model and design a self-attention-based Bi-LSTM model to interpret the model output while keeping the trustable prediction. We consider every feature in a crime sample as particles in a physical system interacting with each other. For example, the number of bookings relates to the number of different level crimes in the record. The mathematical framework of quantum physics provides posthoc interpretations of criminal charge prediction to a certain degree.

The model interpretations shows that our models do not need to rely on personal information to make predictions such as race and age. Our findings underscore eliminating individual information inputs that can cause biased decisions. We summarize our four major contributions as follows:

1) We study four research questions on an individual's future criminal charge prediction and financial predictions. We also introduce Bi-LSTM with attention, focal loss, and oversampling to alleviate the problem of imbalanced data and to interpret feature importance for decision-making.
2) Our Bi-LSTM models with different attention mechanisms and the CVN model provide interpretability for deep learning models. For criminal charge prediction tasks, attention weights and feature-dependent weight in CVN illustrate that demographic characteristics such as race and age are insignificant factors for model outputs. This observation remains consistent when examining financial tasks as well.
3) Our ablation study shows that removing the demographic, which can lead to bias, almost does not influence the prediction results.
4) Our models indicate that a person is more likely to commit the same or a similar level of crime in the near future than a very different level of crime.

II. Background and Related Work

The explainability of the prediction models and understanding the decision-making in highly sensitive areas is of great importance in many domains such as criminal analysis. [12]. Due to the crucial trust-related and ethical issues, the black-box nature of DL models has become one of the primary obstacles to their wide adoption in these areas. In this work, we aim to propose a fair, transparent, and explainable prediction method for these tasks while maintaining sufficient accuracy.

There has been an increasing trend in criminal justice to leverage machine learning. In [13], Yu et al. use Support Vectors Machine (SVM) model, 2-layer feed forward neural network and Naive Bayes model for crime forecasting. Stec et al. [14] and Stalidis et al. [15] use feed forward, convolutional and recurrent-convolutional networks and Luo et al. [16] use recurrent based networks with attention mechanism to analyze law articles. One main characteristic of these high-stake tasks is imbalanced data. There are many ways to address the imbalanced data problem. Astor et al. [17] use oversampling method to analyze crime patterns. Moreover, Lin et al. [18] proposed focal loss to make the model pay more attention to the complex examples. Mulyanto et al. [19] show the effectiveness of focal loss for minority classification. To our knowledge, we are the first to use attentional mechanisms, focal loss, and oversampling techniques in individual criminal charge prediction.

The current methods face challenges regarding interpretability and fairness. In the context of criminal charges, computer software is utilized by courts in the United States to predict future criminal behavior, impacting bail and sentencing decisions [20]. However, concerns have been raised regarding racial bias, with critics suggesting a preference for white defendants. A study [21] in the finance industry emphasizes the need for intelligent algorithms and unbiased decision-making systems that provide accurate and reliable information to fulfill customer demands. It is crucial to understand how decisions are made and whether they are influenced by biased factors such as race and gender. Our evaluation aims to assess the accuracy and fairness of computer software compared to human decision-making, while also exploring the presence of racial bias in predictive algorithms.

III. Task Description and Dataset

A. Crime Prediction

1) Criminal Charge Dataset: A criminal charge is a formal accusation made by a governmental authority asserting that an individual has committed a crime. There are three primary classifications of criminal offenses: felonies (we call these level 1 crimes), including terrorism, murder, kidnapping, treason, and elder abuse; misdemeanors (level 2), such as arson, extortion, threat, bribery, larceny; and infractions (level 3), the least severe crimes, including damaging property, burglary, smuggling, obscenity. The real-life crime prediction data we use contains the criminal charge records in Newark, New Jersey, from 1997 to 2017, with $17,335$ suspect individuals. Each suspect is provided with her/his personal ID, race, and a list of bookings. Each booking has a list of sentences, where each sentence has the age of committing the crime and the individual criminal charge history, such as the NCIC crime code, NCIC category code, and the crime level. The input features also include the number of bookings, age average, number of crimes at levels 1, 2, 3, and so on.

The data statistics of this dataset is shown in Table I. We not only calculated the positive and negative label distribution on the whole data set, we also show the distribution on different features. We observe from Table I that the criminal data is highly imbalanced, with most of the instances having the "No crime" label. This resembles a real-life scenario, where a suspect stops committing a crime more often than continuing to commit crimes. It is technically challenging to handle imbalanced data. A fully connected feed-forward neural network does not work in our experiments since it may classify all samples as negative. For example, suppose there are only 2 positive samples among 100 samples. In that case, the precision is still 98%, although nothing is classified. Moreover, the statistics calculated on different age ranges and races show that these personal characteristics should not be the crucial factors in deciding a person will commit a crime or not.

2) Task: We take the first 18 years as the training set and the last two years as the test set. A suspect with at least two bookings is one sample. The label for each sample is the crime level (or whether it is a specific crime level) during the last booking. Our goal is to predict a person's crime or crime level (1 for most severe and 3 for least severe) given her/his personal information and previous criminal records. We propose a multitude of tasks addressed by Question (Q)1/2/3/4:

"Is a suspect going to commit a crime at level 1/2/3/any?"

Level "any" means any crime of level 1 or 2, or 3. If the label of any of levels 1, 2, or 3 for one person is positive, level any will be positive.

B. Financial Predictions

1) Default Credit Card Payment: The task is to predict whether a customer will have default credit card payment next month. The dataset contains information on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients in Taiwan from April 2005 to September 2005. The number of non-default payment is 6636, the number of default payment is 23364. The protected attributes in this dataset are gender, education level, and age.

2) Fraudulent Transaction Detection: The task is to predict whether a transaction is fraudulent or benign. This synthetic dataset is collected from BankSim which is an agent-based simulator of bank payments based on a sample of aggregated transactional data provided by a bank in Spain. There are 594,643 records in total, 587,443 are normal payments and 7,200 fraudulent transactions. The protected attributes in the dataset are age and gender.

Age/Race		**Crime level** 1		**Crime level** 2		**Crime level** 3	
		Yes	**No**	**Yes**	**No**	**Yes**	**No**
All		6.7	93.3	3.5	96.5	8.7	91.3
Age	≤ 20	16.9	83.1	7.2	92.8	12.8	87.2
	21-30	7.8	92.2	4.3	95.7	8.1	91.9
	31-50	5.9	94.1	2.9	97.1	9.1	90.9
	>50	2.8	97.2	1.4	98.6	7.2	92.8
Race	1	6.8	93.2	3.6	96.4	9.2	90.8
	2	9.8	90.2	2.1	97.9	8.1	91.9
	3	2.6	97.4	2.1	97.9	5.1	94.9
	4	8.6	91.4	3.6	96.4	6.9	93.1
	5	3.1	96.9	0.8	99.2	6.2	93.8

TABLE I: Highly imbalanced data: most bookings are labeled as non-crime. We measure the percentage of the crime (Yes) and non-crime (No) labeled sample number [in $\%$], respectively. The statistics is calculated on the whole dataset (all), and on three different data classifications of the age and race, respectively. To avoid ethical issue, we use 1, 2, 3, 4, and 5 to represent different races.

IV. Methods

A. Alleviating imbalanced data issue with oversampling and focal loss

Although deep neural networks can achieve high accuracy in many domains, the imbalanced datasets suffer from high false-positive or false-negative alarm rates [19], [22]. The models tend to predict all the samples as the majority class to obtain higher accuracy. We can alleviate the imbalance data problem from the data level and algorithm level. Oversampling approach from the data level can balance the dataset by generating new samples in the classes which are under-represented. We over-sample the minority class by picking samples at random with replacement. From the algorithm level, we adopt a cost-sensitive learning cross-entropy with focal loss. It reshapes the typical cross-entropy loss so that the loss can down-weight easy examples and focus more on hard examples. The focal loss is defined as below:

$$FL\left(p\right) = -a\left(1-p\right)^{\gamma}\log\left(p\right), \tag{1}$$

where p is the model's estimated probability for the class with positive label. a balance the importance of positive and negative examples and γ is the modulating factor to make the loss function down-weight easy examples and thus focus training on hard ones.

B. Interpreting Feature Importance with Attention Mechanisms

We provide a certain degree of intrinsic-interpretability using the attention mechanism [23]. In particular, the attention mechanism allows to focus on different parts of the input features to generate a prediction. The attention scores obtained while predicting will indicate the importance of the input features towards the decision and can provide intrinsic interpretability.

We consider three different ways based on the attention mechanism to interpret the contribution of the input features to model predictions. The first two use additive attention with context [24] and the third one uses self-attention.

For additive attention with context, assume the output at each time step in the Bi-LSTM model is $\mathbf{h}_t \in \mathbb{R}^d$, we first calculate the context vector $\mathbf{u}_t$ for every time step with $\mathbf{h}_t$:

$$\mathbf{u}_t = \tanh(\mathbf{W}_h \mathbf{h}_t + \mathbf{b}_t) \left(\mathbf{W}_h \in \mathbb{R}^{d' \times d}, \mathbf{b}_h \in \mathbb{R}^{d'} \right), \quad (2)$$

$\mathbf{W}_h$ is the weight matrix and $\mathbf{b}_t$ is the bias term. Then we calculate the attention weight α_t for each $\mathbf{h}_t$ and sum the weighted context vectors $\{\mathbf{v}_t\}_{t=1,\dots,T}$ up as the input $\mathbf{v}$ of the classification layer:

$$\alpha_t = \frac{\exp(\mathbf{u_t}^T \mathbf{u_t})}{\sum_t \exp(\mathbf{u_t}^T \mathbf{u_t})} \in \mathbb{R} \quad (3)$$

$$\mathbf{v} = \sum_t \alpha_t \mathbf{h}_t \in \mathbb{R}^d. \quad (4)$$

Lastly, we calculate the binary class probability p_a for the input sequence based on $\mathbf{v}$:

$$p_a = \text{Sigmoid}(\mathbf{W}_o \mathbf{v} + \mathbf{b}_o) \left(\mathbf{W}_o \in \mathbb{R}^{1 \times d}, \mathbf{b}_o \in \mathbb{R} \right), \quad (5)$$

where $\mathbf{W}_o$ and $\mathbf{b}_o$ are the parameters for classification layer.

For self attention based model, different from the former model, we add one extra input with symbol CLS at the beginning of the sequence and only use the representation vector $\mathbf{c}$ of CLS for classification:

$$p_{sa} = \text{Sigmoid}(\mathbf{W}_o^{'} \mathbf{c} + \mathbf{b}_o^{'}). \quad (6)$$

p_{sa} is the output probability of a sample. $\mathbf{c}$ is calculated based on all the input features, and $\mathbf{W}_o^{'}$ and $\mathbf{b}_o^{'}$ are the parameters in the classification layer of the model. When calculating $\mathbf{c}$, we use self-attention mechanism. $\mathbf{c}$ is a weighted sum of all value vectors $\{\mathbf{v}_t\}$ for every time step. First, given a sequence of input, we compute the output $\mathcal{H} = \{\mathbf{h}_0, \dots, \mathbf{h}_T\}$ from Bi-LSTM, $\mathbf{h}_t \in \mathcal{H}$ is the representation of the input at time step t and $\mathbf{h}_0$ is of input CLS. After getting the hidden representation $\mathcal{H}$, we compute the query $\mathbf{q}_0$ for $\mathbf{h}_0$ and compute keys $\{\mathbf{k}_t\}$ and values $\{\mathbf{v}_t\}$ for each $\mathbf{h}_t, t = \{0, \dots, T\}$. The transformations are defined as follows:

$$\mathbf{q}_0 = \mathbf{W}_Q \mathbf{h}_0 + \mathbf{b}_Q \left(\mathbf{W}_Q \in \mathbb{R}^{d_s \times d}, \mathbf{b}_Q \in \mathbb{R}^{d_s} \right) \quad (7)$$

$$\mathbf{k}_t = \mathbf{W}_K \mathbf{h}_t + \mathbf{b}_K \left(\mathbf{W}_K \in \mathbb{R}^{d_s \times d}, \mathbf{b}_K \in \mathbb{R}^{d_s} \right) \quad (8)$$

$$\mathbf{v}_t = \mathbf{W}_V \mathbf{h}_t + \mathbf{b}_V \left(\mathbf{W}_V \in \mathbb{R}^{d_s \times d}, \mathbf{b}_V \in \mathbb{R}^{d_s} \right). \quad (9)$$

Having the query vector $\mathbf{q}_0$ for $\mathbf{h}_0$ and key vectors $\{\mathbf{k}_t\}$ for $t = 1, \dots, T$, we can get the attention weight $\alpha_{0,t}$ assigned to each value vector of $\mathbf{v}_t$ with the equation below:

$$\alpha_{0,t} = \text{Softmax}\left(\mathbf{q}_0^T \mathbf{k}_t \right) \in \mathbb{R} \quad (10)$$

Then we gather all value vectors $\{\mathbf{v}_t\}$ of $\mathbf{h}_t \in \mathcal{H}$ based on the attention weights $\alpha_{0,t}$, $t = 0, \dots, T$ to get representation $\mathbf{c}$ of CLS in Equation 6:

$$\mathbf{c} = \sum_{t=0}^{T} \alpha_{0,t} \mathbf{v}_t \quad (11)$$

After prediction, we focus on the interpretation of these two attention-based models. Firstly, we use the attention weight α_t of additive attention to representing the importance of each input feature, which is the weight-based analysis. The larger α_t is, the higher contribution this feature makes to the prediction. Secondly, we consider the norm-based attention analysis for prediction interpreting [25]. Instead of directly using the attention weight to evaluate the importance of the feature for prediction, we use the norm w_t defined below to interpret the significance of each input feature:

$$w_t := \|\alpha_t \mathbf{h}_t\| \quad (12)$$

The third one to interpret the model prediction is the weight-based analysis with self-attention. Each attention weight $\alpha_{0,t}$, $t = 0, \dots, T$ in self-attention shows the importance of input feature at t on the model prediction. Similar to the first analysis, we can use these attention weights to interpret the model outputs.

C. Increasing Model Transparency with Quantum-inspired Complex Valued Networks

As discussed before, the attention weights can fail to identify the important features of the model's final decision. Therefore, we adopt another quantum-theoretical framework for the model interpretation with explicit physical meanings. We adopt the model of [26] from the matching task to a binary classification task. We first model the crime records with the mathematical framework of quantum physics and then introduce the interpretability of the model with the physical meaning. The model architecture is shown in Fig. 1.

We consider a criminal record as a physical system. The features of a criminal record can be treated as particles, and each feature is composed of a set of orthogonal basis states in superposition. For example, the basis states can include vectors representing a specific crime level in a random year and other meanings. A feature, for instance, the number of bookings will be composed of all basis states. We use a complex-valued vector $\vec{w_j}$ to represent each feature w_j. Each feature vector is normalized into a superposition state $|w_j\rangle$:

$$|w_j\rangle = \frac{\vec{w_j}}{\|\vec{w_j}\|}, \quad (13)$$

$\|\vec{w_j}\|$ denotes the L_2-norm of $\vec{w_j}$, which is used to calculate the relative weight of the j-th feature in a crime record. Viewing a crime record as a mixed system composed of individual features, the representation of a crime record is computed as follows, where ρ is the mixture density matrix:

$$\rho = \sum_{i=1}^{n} p(w_i) |w_i\rangle \langle w_i| \quad (14)$$

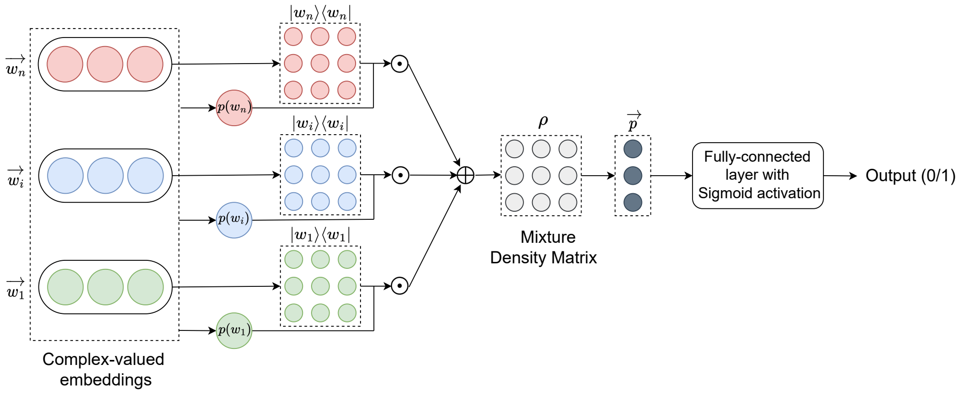

$\overrightarrow{w_n}$
$\overrightarrow{w_i}$
$\overrightarrow{w_1}$
$|w_n\rangle\langle w_n|$
$|w_i\rangle\langle w_i|$
$|w_1\rangle\langle w_1|$
$p(w_n)$
$p(w_i)$
$p(w_1)$
ρ
$\vec{p}$
Complex-valued embeddings
Mixture Density Matrix
Fully-connected layer with Sigmoid activation
Output (0/1)

Fig. 1: Quantum-Inspired Complex-valued Network.

Here, n is the total number of features of a record and $p(w_i)$ is the soft-max normalized feature dependent weight: $p(w_i) = \frac{e^{\|\vec{w_i}\|}}{\sum_{i=1}^{n} e^{\|\vec{w_i}\|}}$. Then ρ is projected using K measurement matrices to get a probability vector $\vec{p}$ of dimension K. A fully connected layer with sigmoid activation is applied on $\vec{p}$ for the binary classification. In this system, the value of $\|\vec{w_j}\|$ represent the importance of the j-th feature, which can be used to interpret the model prediction.

V. Experimental Setup

Our models are based on Bi-LSTM with attentions [27], where each hidden layer has 64 nodes, i.e. $d = 64$ with or without attention layers. The dropout parameter is set to 0.2. The parameters are randomly initialized and updated using the Adam optimizer [28] with a learning rate of 0.0005. We set 50 epochs for training. For class prediction, we use Sigmoid activation to get the output probability with a threshold of 0.5. If the output probability is greater than 0.5, the sample is considered positive, otherwise, it is negative.

The attention mechanisms we used upon the Bi-LSTM are additive attention and self attention. For additive attention, the dimension of the context vector d' is 100, and for self-attention, the dimensions d_s for query, key, and value are the same, which are all set to 256. For the CVN model, we set the complex-valued embedding dimension as 100, and the number of measurement matrices K is also 100.

When addressing the imbalanced data problem, we use two approaches from the data and algorithm levels. From the data level, we over-sample the minority samples and make the sample number of minority classes equivalent to that of the majority class. For the algorithm-level based method, we adopt focal loss with the hyper-parameters $\gamma = 2$ and $a = 0.5$.

VI. Prediction Results

A. Crime charge prediction

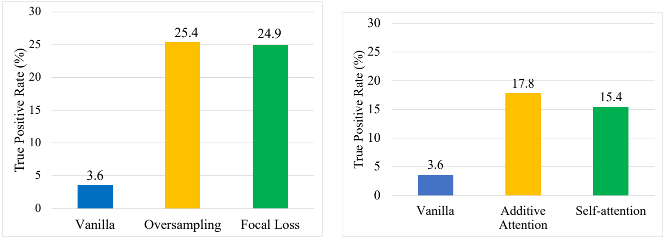
For the crime prediction, We consider that different features should contribute differently to the model prediction output. For example, the criminal history might be a much more important factor in deciding whether a person will have a new crime or not. Therefore, we add the attention mechanism to Bi-LSTM. We adopt two different attention mechanisms. One is additive attention, and the other one is self-attention. The prediction results of Bi-LSTM with these two attentions are shown in Table II. The Bi-LSTM models with different attention methods achieve similarly high accuracy, precision-recall, and F1 scores, thus they have comparable performance. Comparing the prediction results of the CVN model with attention-based Bi-LSTM models, CVN has the best performance in general on all metrics and all levels.

Model	level	Acc.	Prec.	Recall	F1
Bi-LSTM	1	92.3	91.4	92.3	91.8
	2	95.6	93.6	95.6	94.6
	3	87.4	87.8	87.4	88.2
	Any	87.9	87.5	87.9	87.7
Bi-LSTM with additive attention	1	94.4	92.8	94.4	92.9
	2	93.9	94.2	93.9	94.1
	3	87.8	88.0	87.8	88.1
	Any	87.4	86.8	87.4	87.0
Bi-LSTM with self attention	1	94.4	92.7	94.4	92.8
	2	94.1	94.2	94.1	94.2
	3	87.8	88.1	87.8	88.0
	Any	86.3	88.9	86.3	87.1
CVN	1	94.0	91.6	94.0	92.1
	2	96.0	94.1	96.0	94.9
	3	89.1	88.0	89.1	88.5
	Any	88.1	87.3	88.1	87.5

TABLE II: Crime charge prediction using Bi-LSTM models with and without attention, as well as the CVN model on different levels.

B. Addressing imbalanced data issue

For the crime prediction tasks, according to Table I, we notice that crime level 2 has a quite low proportion of positive samples, which is only about 3%. To further investigate whether the models will suffer from the imbalance data issue on crime level 2 prediction, we plot the true positive rate of the prediction of level-2 crime for Bi-LSTM. From Fig. 2a and Table II we can see that, although the accuracy is pretty high for level-2 crime, the models' predictions of Bi-LSTM have very low true positive rates, which means the models tend to predict the positive samples as negative. There are 551



(a)

(b)

Fig. 2: True positive rate of Bi-LSTM models with different approaches that can alleviate imbalanced data .

Model	Acc.	Prec.	Recall	F1
Bi-LSTM	95.6	93.6	95.6	94.6
Bi-LSTM with oversampling	92.7	94.4	92.7	93.5
Bi-LSTM with focal loss	93.2	94.4	93.2	93.8

TABLE III: Bi-LSTM, Bi-LSTM with oversampling, and Bi-LSTM with focal loss results for crime level 2 prediction.

	Crime level (L)	$L = 1$	$L = 2$	$L = 3$
Correlation	$P(\text{level } 1 \mid L)$	0.53	0.37	0.25
	$P(\text{level } 2 \mid L)$	0.15	0.33	0.19
	$P(\text{level } 3 \mid L)$	0.32	0.30	0.56

TABLE IV: Crime level transitions. Given L as the label, the probability/weights of each lowest (most severe crime) level in history.

positive samples in the test dataset for crime level 2 while the Bi-LSTM model only predicts 20 positive samples correctly.

To address this imbalanced data issue, we adopt two approaches. The first one is oversampling the minority class to make the dataset balanced, and the second one is to assign higher weight in the loss function to the hard predict samples with focal loss [18]. Table III also shows the prediction results on the four evaluation metrics Bi-LSTM with oversampling and focal loss. We plot the true positive rate in Fig. 2a. A higher true positive rate means more positive examples are classified correctly. Compared with the vanilla Bi-LSTM, Bi-LSTM with oversampling and focal loss can alleviate the low true positive prediction rate problem, which results from the too-small proportion of the positive samples in the dataset. In addition, we also show the true positive rate on level 2 crime prediction for the vanilla Bi-LSTM model and the models with two attention methods in Fig. 2b. We notice that Bi-LSTM with attention-mechanism can also improve the true positive rate to some extent. With attention, the model can focus more on the features that are more important to the prediction and improve the true positive prediction rate.

In conclusion, oversampling the minority class, using focal loss instead of traditional cross-entropy loss, and adopting attention-mechanism can alleviate the imbalanced data problem in the level 2 crime prediction task.

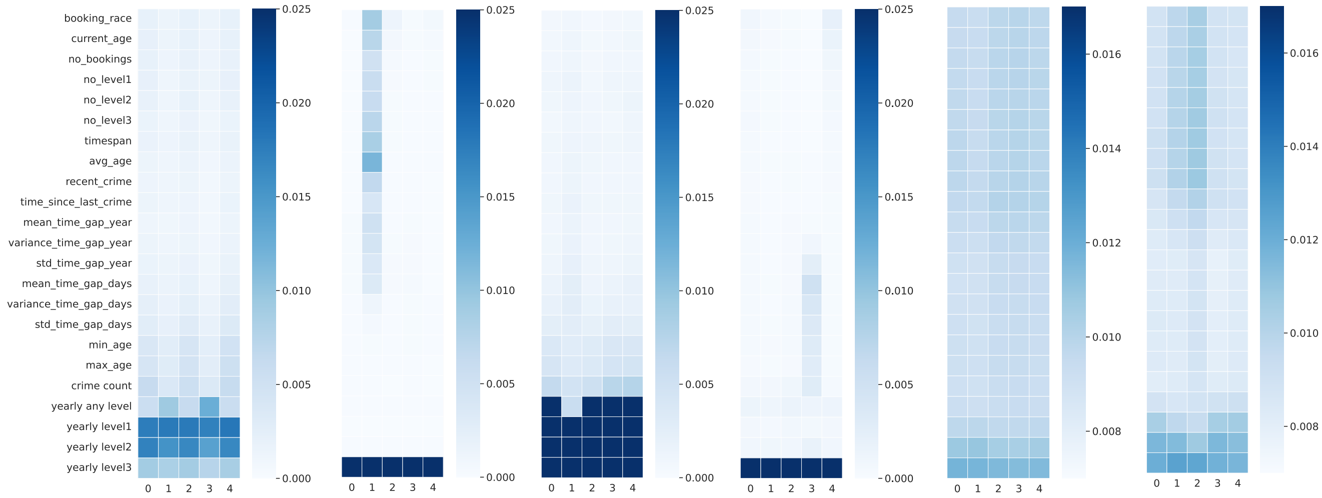
C. Interpreting Results

First, we analyze the interpretability of the Bi-LSTM model for crime prediction tasks and take level 2 and level 3 crime predictions as examples. We group some features of the same type and average the attention weights for visualization. For example, we group all years' criminal records of level 1 and name the group as yearly level1. The visualization of attention weights α_t, the norm of $\alpha_t \mathbf{h}_t$ for additive attention and attention weights $\alpha_{0,t}$ for self attention on Bi-LSTM model are shown in Fig. 3. We randomly selected five test samples in level 2, and level 3 crime charge perdition.

We also show the feature weights in Quantum-Inspired CVN in Fig. 4. Different from the attention-based framework, more features of the input have relatively larger weights in Quantum-Inspired CVN. For example, only the number of yearly crimes (yearly level1, level2, and level3) contribute most to the prediction in the attention-based models, and the weights of other features are almost zeros. However, features such as recent crime and the number of each level of crime have larger importance in Quantum-Inspired CVN, and few features have weights close to zeros. For example, in Fig. 4, the features number of level1 booking, average, time span, and information about gap times contribute more to the final prediction. For level any, the number of level 1, 2, and 3 bookings, average age, and recent crime time have larger importance. Therefore, we infer that because the decision of Quantum-Inspired CVN is made by more features rather than only a few, it can achieve a better performance than attention-based models.

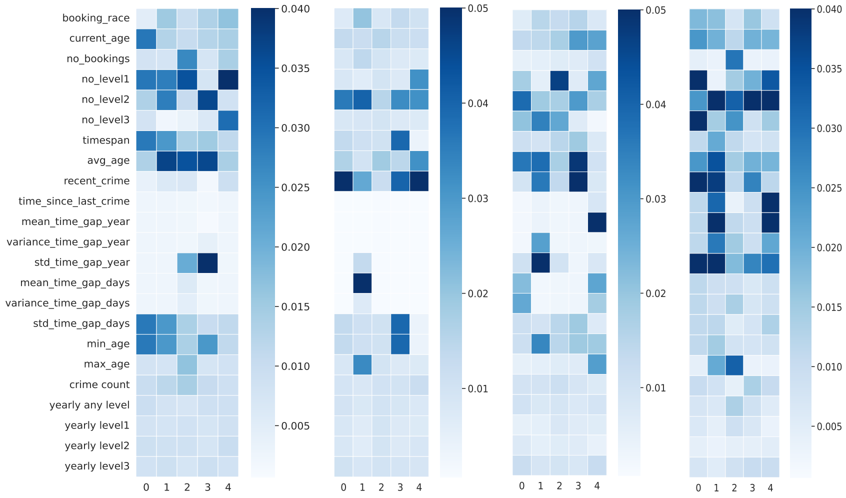
Importantly, Fig. 3 and Fig. 4 show that personal information such as race and age contribute very little towards the prediction. Instead, an individual's criminal history features greatly impact the prediction, such as "yearly crime records", "number of different level bookings, and "recent crime time". This indicates that personal information features do not explicitly contribute to our algorithmic decisions. Note that our work only focuses on the feature importance explicitly used in our model. Discussions on any bias made by humans in the previous charge decisions [29], [30] are out of the scope of this paper.

In our correlation study on the transitions of a suspect from one crime level to another, presented in Table IV, we find that crime level features are essential to predict the level of a new crime. With the visualization of the attention weights, we can also observe a similar phenomenon. The history of a specific crime level significantly influences the prediction of the same level. For example, in Fig. 3a the additive attention weights interpret that the history of level 1 and level 2 crime makes the most contribution to the prediction of level 2 crime. When predicting the level 3 crime, level 3 crime history makes the most contribution, according to Fig. 3b and Fig. 3d. Both our feature contribution results, obtained from our Bi-LSTM model with attention mechanisms, and the correlation results indicate that a suspect is more likely to commit a crime at the same or a similar level again, rather than at a significantly



(a) (b) (c) (d) (e) (f)

Fig. 3: Different weights of the input features for crime charge prediction. (a) and (b) are additive attention weights of level 2 and level 3 crime charge prediction. (c) and (d) are norm weights $\|\alpha_t \mathbf{h}_t\|$ of level 2 and level 3 crime charge prediction. (e) and (f) are self-attention weights of level 2 and level 3 crime charge prediction. The x-axis are five randomly selected samples from the test dataset. The y-axis presents the names of the features.



(a) level 1 (b) level 2 (c) level 3 (d) level any

Fig. 4: Feature weights of the input in Quantum-Inspired CVN. The x-axis are five randomly selected samples from the test dataset. The y-axis present the name of the features.

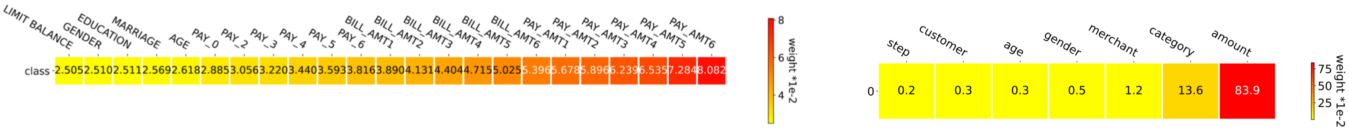
higher or lower level.

We also perform interpretation study on the Bi-LSTM models on the two financial tasks. We observe similar patterns from these two models. Fig. 5 shows the feature importance for both default credit card payment prediction task and fraudulent transaction detection task. As shown in the figure, for the default payment prediction, the protected demographic features, i.e., gender, education level, and marriage status, do not contribute a lot to the model's predictions, which means the model does not rely on the demographic features to make decisions. Similarly, for the fraudulent detection task, the sensitive features age and gender are not important to the model's decision making.

D. Fairness Study: Prediction of removing the personal information

Sensitive attributes such as race, age, and gender can induce bias toward the model and might incur ethical issues. Based on the model interpretation that these sensitive attributes are not the crucial factors for the model prediction, we assume that removing these features will not significantly influence the model prediction.

In the fairness study, we first remove the race and age in crime charge prediction tasks. Table V displays the prediction



(a)

(b)

Fig. 5: Attention scores of Bi-LSTM models on two financial datasets. (a) shows the feature importance of the default credit card payment prediction task, and (b) shows that of the fraudulent transaction detection task.

task		Acc.	Prec.	Recall	F1
Crime charge	level 1	94.4	93.0	94.4	93.1
	level 2	95.0	94.0	95.0	94.5
	level 3	89.9	87.8	89.9	88.7
	level any	86.7	87.4	86.7	87.0

TABLE V: Prediction results after removing the personal attributes on crime charge using Bi-LSTM model.

task		Acc.	Prec.	Recall	F1
Crime charge	level 1	92.8	92.2	92.8	92.5
	level 2	94.7	94.3	94.7	94.5
	level 3	90.7	86.3	90.7	87.8
	level any	88.4	88.0	88.4	88.1

TABLE VI: Prediction results after removing the personal attributes on crime charge using CVN model.

results of Bi-LSTM with additive attention on the crime charge prediction tasks after removing personal attributes. Similarly, Table VI presents the prediction results of CVN on the same tasks with personal attributes removed. Comparing Table V and Table VI with Table II, we can see that, removing the sensitive features will not influence the model performance. Similarly, we remove the sensitive features from the credit card default payment dataset. Table VII shows the metrics before and after the sensitive personal features are removed. Similarly, removing the sensitive features do not influence the model performance significantly.

These personal features might raise concerns about whether the models are fair, and by excluding these features, we achieve more fair models with the same accuracy as the baseline models. Therefore, with the interpretation of the model and the features, we can conclude that the sensitive features do not contribute a lot to the model's decision making process, and thus we can achieve both accuracy and fairness with this approach.

VII. Interpretability/Explainability Evaluation

To quantify the interpretability of each of the proposed methods, we look at the attention scores for each of the input features for a given input sample. Fig. 3 shows the attention weights for different models and different attention mechanisms for five randomly selected test samples. For example, Fig. 3a shows that yearly level 1 and yearly level 2 features contribute the most for level 2 crime prediction when an additive attention mechanism was used. Similarly, Fig. 3f shows that along with yearly level features, other features including the number of crimes for each level and the time between crimes, are also important when making the prediction. In general, these attention weights quantitatively show how crime-related features are more important to the prediction compared to the sensitive features, including the person's race or age.

Features	**Recall**	**Precision**	**F-1 score**
All Features	0.807	0.787	0.789
Without sensitive features	0.805	0.784	0.785

TABLE VII: Bi-LSTM performance on default payment dataset. The metrics shown above are weighted average of both classes.

Experiment	Prediction	Percentage change
Bi-LSTM with additive attention	True	0.724
	False	0.057

TABLE VIII: Percentage prediction changed when pruning attention

To further analyze the influence of each of the features, we use the attention zeroing method. For each test input, we randomly pruned 10% of the attention layer neurons. Then we calculate the total number of predictions that changed. We experiment with both the LSTM with additive attention and Bi-LSTM with additive attention for any level crime prediction task on randomly selected 100 test samples. The results are mentioned in Table VIII. This shows that Bi-LSTM is heavily dependent on the neurons in the attention layer.

VIII. Conclusion

Our work introduces trustable prediction methods with high precision, high recall, and model interpretability for two high-stake areas: crime prediction and finance. We adopt two approaches to address the imbalanced data problems and introduce three attention-based methods to interpret model predictions. We show that deep learning networks with model interpretability can be a part of criminal justice assistant systems and financial systems as long as model transparency and accuracy are taken care of. Perhaps most importantly, we draw attention to the erroneous assumption and demonstrate that social features such as race and age are statistically insignificant to influence our model prediction, even though data may introduce bias. Our approach promotes fairness without sacrificing accuracy, advancing fair deep learning in high-stakes domains utilizing big data and cloud computing.

References

[1] J. Huttunen, J. Jauhiainen, L. Lehti, A. Nylund, M. Martikainen, and O. M. Lehner, "Big data, cloud computing and data science applications in finance and accounting," *ACRN Journal of Finance and Risk Perspectives*, vol. 8, pp. 16–30, 2019.

[2] A. Abdelaziz, M. Elhoseny, A. S. Salama, and A. Riad, "A machine learning model for improving healthcare services on cloud computing environment," *Measurement*, vol. 119, pp. 117–128, 2018.

[3] J. Weidong, "The change of judicial power in china in the era of artificial intelligence," *Asian Journal of Law and Society*, vol. 7, no. 3, pp. 515–530, 2020.

[4] P. P. Angelov, E. A. Soares, R. Jiang, N. I. Arnold, and P. M. Atkinson, "Explainable artificial intelligence: an analytical review," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 11, no. 5, p. e1424, 2021.

[5] A. Hanif, "Towards explainable artificial intelligence in banking and financial services," *arXiv preprint arXiv:2112.08441*, 2021.

[6] S. Khan and T. Yairi, "A review on the application of deep learning in system health management," *Mechanical Systems and Signal Processing*, vol. 107, pp. 241–265, 2018.

[7] Y. Zhang, P. Tiňo, A. Leonardis, and K. Tang, "A survey on neural network interpretability," *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2021.

[8] C. Huang, J. Zhang, Y. Zheng, and N. V. Chawla, "Deepcrime: Attentive hierarchical recurrent networks for crime prediction," in *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, 2018, pp. 1423–1432.

[9] S. Hossain, A. Abtahee, I. Kashem, M. M. Hoque, and I. H. Sarker, "Crime prediction using spatio-temporal data," in *International Conference on Computing Science, Communication and Security*. Springer, 2020, pp. 277–289.

[10] S. Kim, S. Ku, W. Chang, and J. W. Song, "Predicting the direction of us stock prices using effective transfer entropy and machine learning techniques," *IEEE Access*, vol. 8, pp. 111 660–111 682, 2020.

[11] A. Moghar and M. Hamiche, "Stock market prediction using lstm recurrent neural network," *Procedia Computer Science*, vol. 170, pp. 1168–1173, 2020.

[12] N. Burkart and M. F. Huber, "A survey on the explainability of supervised machine learning," *Journal of Artificial Intelligence Research*, vol. 70, pp. 245–317, 2021.

[13] C. Yu, M. W. Ward, M. Morabito, and W. Ding, "Crime forecasting using data mining techniques," in *2011 IEEE 11th International Conference on Data Mining Workshops*, Dec 2011, pp. 779–786.

[14] A. Stec and D. Klabjan, "Forecasting crime with deep learning," *arXiv preprint arXiv:1806.01486*, 2018.

[15] P. Stalidis, T. Semertzidis, and P. Daras, "Examining deep learning architectures for crime classification and prediction," *arXiv preprint arXiv:1812.00602*, 2018.

[16] B. Luo, Y. Feng, J. Xu, X. Zhang, and D. Zhao, "Learning to predict charges for criminal cases with legal basis," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 2727–2736.

[17] J. Asor, F. Balahadia, G. M. Catedrilla, and M. V. Villarica, "Building model for crime pattern analysis through machine learning using predictive analytics," *International Journal of Science, Technology, Engineering and Mathematics*, vol. 2, no. 1, pp. 61–73, 2022.

[18] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[19] M. Mulyanto, M. Faisal, S. W. Prakosa, and J.-S. Leu, "Effectiveness of focal loss for minority classification in network intrusion detection systems," *Symmetry*, vol. 13, no. 1, p. 4, 2020.

[20] J. Dressel and H. Farid, "The dangers of risk prediction in the criminal justice system," *MIT Case Studies in Social and Ethical Responsibilities of Computing*, 2021.

[21] P. Bracke, A. Datta, C. Jung, and S. Sen, "Machine learning explainability in finance: an application to default risk analysis," 2019.

[22] E. Hodo, X. Bellekens, A. Hamilton, C. Tachtatzis, and R. Atkinson, "Shallow and deep networks intrusion detection system: A taxonomy and survey," *arXiv preprint arXiv:1701.02145*, 2017.

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.

[24] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[25] G. Kobayashi, T. Kuribayashi, S. Yokoi, and K. Inui, "Attention is not only a weight: Analyzing transformers with vector norms," *arXiv preprint arXiv:2004.10102*, 2020.

[26] Q. Li, B. Wang, and M. Melucci, "CNM: An interpretable complex-valued network for matching," *arXiv preprint arXiv:1904.05298*, 2019.

[27] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[28] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[29] P. J. Brantingham, M. Valasik, and G. O. Mohler, "Does predictive policing lead to biased arrests? results from a randomized controlled trial," *Statistics and public policy*, vol. 5, no. 1, pp. 1–6, 2018.

[30] D. Arnold, W. Dobbie, and C. S. Yang, "Racial bias in bail decisions," *The Quarterly Journal of Economics*, vol. 133, no. 4, pp. 1885–1932, 2018.